DOCUMENT RESUME

ED 423 240 TM 028 278

AUTHOR Krass, Iosif A.
TITLE Application of Direct Optimization for Item Calibration in Computerized Adaptive Testing.
PUB DATE 1998-00-00
NOTE 35p.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; Algorithms; *Computer Assisted Testing; Estimation (Mathematics); Online Systems; *Test Items
IDENTIFIERS *Calibration; Estimation; Item Parameters; Likelihood Function Estimation; *Optimization

ABSTRACT

In the process of item calibration for a computerized adaptive test (CAT), many well-established calibrating packages show weakness in the estimation of item parameters. This paper introduces an on-line calibration algorithm based on the convexity of likelihood functions. This package consists of: (1) an algorithm that estimates examinee ability and (2) an algorithm that estimates the parameters for a new item that is seeded into the CAT test. The performance of the new package is comparable to BilogMG, and in some cases exceeds it. The new algorithm belongs to the class of Direct Maximization Aposteriori algorithms. (Contains 9 figures, 2 tables, and 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 423 240

Running head: ITEM CALIBRATION IN COMPUTERIZED ADAPTIVE TESTING

Application of Direct Optimization for Item Calibration in Computerized Adaptive Testing

Iosif A. Krass

Defense Manpower Data Center

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Iosif Krass

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The author would like to thank Daniel Segall, Robert Holmes, and Bruce Bloxom for their supporting comments and helpful suggestions.

Requests for reprints should be sent to Iosif A. Krass, Defense Manpower Data Center, 400 Gigling Road, Seaside, CA 93955-6771. E-mail: krassia@pentagon.mil

TM028278

Abstract

In the process of item calibration for a CAT test, many well-established calibrating packages show weakness in the estimation of item parameters. This paper introduces an on-line calibration algorithm based on the convexity of likelihood functions. This package consists of: (a) an algorithm that estimates examinee ability, and (b) an algorithm that estimates the parameters for a new item that is seeded into the CAT test. The performance of the new package is comparable with BilogMG, and in some cases exceeds it.

Key Words: computerized adaptive testing, CAT, item calibration, item parameters, maximization of likelihood, log-likelihood function, precision, BilogMG, DMAP, ICCs, multi-dimensional test, convexity.

Application of Direct Optimization for Item Calibration in Computerized Adaptive Testing

1. Introduction

The problem of item calibration--estimation item parameters when the model of responses is fixed--is very old and has been well discussed in the psychometric literature (e.g., Bock & Aitkin, 1981; Thissen & Steinberg, 1984; Samejima, 1969; Levine, 1984). There are a few packages available, particularly BilogMG, which are designed to do the job of calibration (Bilog 3, 1990; Multilog, 1988). However, nearly all available packages and algorithms are designed to use results of tests given in the classical paper-and-pencil mode. Typically, test results from a computer-adaptive-testing mode very often contradict assumptions underlying typical calibration packages, and application of those packages generally leads to large biases and standard errors in item-parameter estimates for a computer-adaptive test.

These constraints have recently been addressed in work with the Armed Services Vocational Aptitude Battery (ASVAB) computerized adaptive testing mode (CAT) which uses a seeded-item design (Segall, Moreno, Bloxom, & Hetter, 1997) to get parameters for new items. The CAT-ASVAB seeded-item program allows access to an unbiased examinee population and adds little additional cost to the ongoing operational testing. However, in CAT testing, the matrix of examinee-by-item responses is rather sparce, in comparison with the classical paper-and-pencil test. The CAT tests are rather short (at most 15 items) because of computerized-adaptation to each examinee, and the examinee population is sometimes considerably different

from a normal-normal population (i. e. ability distribution of population is normal with mean zero and standard error one). Moreover, there is an obvious violation of the single dimension assumption for at least one test, the General Science test (Zimowski, 1987). Therefore, we have developed an algorithm based on likelihood optimization, which is a parametric algorithm type of EM (McLachlan, 1997) and is not marginal; so, it belongs (Baker, 1992) to the class of Direct Maximization Aposteriori algorithms (DMAP). In this paper we will describe the new algorithm and compare it with adjusted BilogMG, the most widely used parametric calibration package.

The DMAP algorithm begins by estimating examinee ability based on the test results (Krass, 1997). Utilizing this estimate, it then estimates the 3PL parameters of the seeded item. Next, DMAP re-estimates examinee ability and continues this process to convergence with the required precision. Thus, we see that DMAP, as a usual calibrating algorithm, is an algorithm of the EM type. In this paper we will describe estimating examinee ability by DMAP and then estimating seeded item parameters by DMAP, and we will present some simulation results to compare the performances of DMAP and BilogMG.

2. Estimation of Examinee Ability

Let our test consist of I items, with Item Characteristic Curve (ICC) $P_i(\theta)$, $i = 1,\cdots,I$ being 3PL ICC, i.e.,

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp(l_i(\theta))}, \tag{1}$$

where $l_i(\theta) = -D \cdot a_i \cdot (\theta - b_i)$, and a_i, b_i, c_i; $i = 1, \cdots, I$ the item discriminating, difficulty, and guessing indexes, correspondingly; θ is the latent ability of an examinee and $D = 1.7$ is a scaling constant (Lord, 1977). We assume that examinee ability $\theta \in [\theta_{\min}, \theta_{\max}]$, which means that the optimization described below should be done as a constrained optimization (a feature which cannot be done with an internal algorithm type such as Newton-Raphson). Typically, in CAT-ASVAB we have agreement $\theta_{\min} = -3.0$ and $\theta_{\max} = +3.0$. Let our examinee get a sequence $\{i_1, i_2, \cdots, i_k\}$ of items generated by CAT, where $k \le K$, and K is the length of the CAT-ASVAB test (usually $10 \le K \le 15$). Remember, the CAT-ASVAB test is totally driven by an information table based on an item pool with a rather large exposure control factor (about 0.7) (Hetter & Sympson, 1997). CAT-ASVAB items are multiple-choice items, so the examinee produces a dichotomous answer sequence $\bar{u}_k = \{u_1, u_2, \cdots, u_K\}$. Then, his or her likelihood function after the first k items of the test is:

$$L(\bar{u}_k, \theta) = g(\theta) \cdot \prod_{i=1}^{k} P_i(\theta)^{u_i} \cdot Q_i(\theta)^{(1-u_i)} \qquad (2)$$

where $Q_i(\theta) = 1 - P_i(\theta)$ and $g(\theta)$ is the density of prior ability distribution in the population of examinees. The value $\bar{\theta}_k$ which maximizes likelihood

$$L(\bar{u}_k, \bar{\theta}_k) = \max_{\theta \in [\theta_{\min}, \theta_{\max}]} L(\bar{u}_k, \theta)$$

is considered to be the best estimator of the examinee's ability after the first k items of the test. As usual, we assume that prior ability distribution is normal $N(\mu, \sigma)$, i.e.,

$g(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(\theta-\mu)^2}{2\cdot\sigma^2})$, where μ and σ are the mean and SE of prior distribution.

Typically, we begin from normal-normal prior $N(0,1)$ and then tune μ and σ to get better convergence. If we begin from $N(0,1)$, to get the maximizing $\bar{\theta}_k$ we consider log-likelihood which has its derivative due to (1) as:

$$\frac{d}{d\theta}\log(L(\bar{u}_k,\theta)) = -\theta + \sum_{i=1}^{k}(\frac{u_i}{P_i(\theta)} - \frac{1-u_i}{1-P_i(\theta)})\cdot\frac{d}{d\theta}P_i(\theta) = -\theta + \sum_{i=1}^{k} R_i(\theta)$$

where

$$R_i(\theta) = \begin{cases} \dfrac{(1-c_i)\cdot\exp(l_i(\theta))\cdot D\cdot a_i}{(1+\exp(l_i(\theta)))\cdot(1+c_i\cdot\exp(l_i(\theta)))}; & for\ u_i = 1 \\ \dfrac{-D\cdot a_i}{(1+\exp(l_i(\theta)))}; & for\ u_i = 0 \end{cases} \qquad (3)$$

To find a zero of log-likelihood derivative, in the case when the log-likelihood maximum is reached inside of domain segment $[\theta_{min},\theta_{max}]$, we must solve the "fixed-point" problem for function $\sum_{i=1}^{k} R_i(\theta)$, i.e., find a solution of the equation:

$$\theta = \sum_{i=1}^{k} R_i(\theta) \qquad (4)$$

Solution of this type of equation is heavily studied in computational mathematics literature (Blum, 1972; Ramsay, 1975), but the fastest solution can be reached in the case of monotone functions $R_i(\theta)$ which we have here. From (3) it follows that, in the case of $u_i = 1$, we have $R_i(\theta) > 0$, and $R_i(\theta) \to 0$ if $\theta \to \pm\infty$. On the other hand, in the case of $u_i = 0$, the function

$R_i(\theta) < 0$, and $R_i(\theta) \to 0$ if $\theta \to -\infty$, and $R_i(\theta) \to -D \cdot a_i$ if $\theta \to +\infty$. Thus,

$\theta > \sum_{j=1}^{k} R_{i_j}(\theta)$ for $\theta = \theta_{\max}$, and $\theta < \sum_{j=1}^{k} R_{i_j}(\theta)$ for $\theta = \theta_{\min}$, if $\theta_{\max}$ is large enough and

$\theta_{\min}$ is small enough. Therefore, depending on whether the answer is right or wrong, the first zero of (3), which defines the DMAP estimation of examinee ability after the first item administered by CAT, can be found by dichotomy from the "right side" if the answer is correct, or "left side" in the opposite case. Under right side, we mean beginning the process of checking if the inequality

$$\theta_{\max} \le \sum_{i=1}^{k} R_i(\theta_{\max}) \tag{5}$$

holds. From (5) it follows that in the case, when (5) holds, the derivative of $\frac{d \log(L(u_k, \theta))}{d\theta}$ is negative in all our domains, so the maximizing latent ability $\overline{\theta}_1 = \theta_{\min}$; in this case the process can be continued to the next item. If the above inequality is not true, we check the left side condition $\theta_{\min} \ge \sum_{i=1}^{k} R_i(\theta_{\min})$ to see if maximization is reached on the right border of the domain. After checking borders we are sure that at least one solution of (4) is inside the segment $[\theta_{\min}, \theta_{\max}]$, and it can be found by the following dichotomy process: Let $\widetilde{\theta}_{\min} = \theta_{\min}$ and $\widetilde{\theta}_{\max} = \theta_{\max}$ define $\widetilde{\theta} = \widetilde{\theta}_{\min} + 0.5 \cdot (\widetilde{\theta}_{\max} - \widetilde{\theta}_{\min})$. If $\widetilde{\theta} < \sum_{j=1}^{k} R_{i_j}(\widetilde{\theta})$, then $\widetilde{\theta}_{\min} = \widetilde{\theta}$ and $\widetilde{\theta}_{\max} = \widetilde{\theta}$ in the case of opposite inequality. The process continues until $\widetilde{\theta}_{\max} - \widetilde{\theta}_{\min} > \delta$, where δ is a

given precision of computation. The algorithm converges with speed $\frac{1}{2^n}$, where is n is the number of iterations.

As it is shown by Samejima (1973), the log-likelihood function (2) is not, generally speaking, uni-modal, so (4) can have more than one solution, but the second solution is usually out of the border of the "normal" domain. In the case of our algorithm, even though it is designed to hunt for more than one solution of (4), after more than 1,000,000 applications of the algorithm to the simulated or real life test situation, we were not able to find a second solution of (4) in the considered domain [-3.0, +3.0].

From the properties of (3) it follows, independently of the first answer, if the answer on the second item is correct, the root of the equation (4) will be moved to the right, and it can be found by dichotomy beginning from the right side. If the answer on the second item in the sequence is wrong, the root of (4) will be moved to the left, and it can be found by dichotomy from the left side. This phenomena is due to the property $R_i(\theta) > 0$ in the case of a correct answer, and $R_i(\theta) < 0$ in the case of a wrong answer. This phenomena reduces the domain of searching of maximizing likelihood ability while the test is developing adaptively.

In Figure 1, we present the case of a test where the first item is answered correctly and the second wrongly. The darker curve corresponds to the function $R_1(\theta)$ for the first correct answer, and the lighter curve corresponds to the summation $R_1(\theta) + R_2(\theta)$ for the first two items when the first was answered correctly and the second wrongly. The intersection of the straight line and the graph of the function $R_1(\theta) + R_2(\theta)$ gives the DMAP estimation of theta for the test length of two.

(Figure 1 about here.)

In the current CAT-ASVAB, the Owen-Baysian algorithm (Owen, 1975) is applied to estimate ability of the examinee "on-the-fly," and the Baysian-Modal (Segall, et al., 1997) algorithm is applied to the total test sequence to make the final tuning in ability examinee estimation. The above described DMAP algorithm requires a little bit more computer time (about 1.5 more), but it gives more precision in the estimation of examinee ability in the densest part of the ability distribution.

The results of a simulation for 3,000 examinees for Arithmetic Reasoning in CAT-ASVAB Form 1, where the size of the item pool is equal to $I = 94$, is shown in Figures 2 and 3. In this simulation experiment, we took 3,000 normal-ability-distributed examinees and "recovered" their known "true" ability by standard Baysian methods (Figure 2) and by the DMAP algorithm (Figure 3).

(Figures 2 and 3 about here.)

As we can see, DMAP has about the same precision (in the sense of SE or maximum–minimum deviation) as a standard Baysian algorithm for $\theta \geq -1.85$ but does better than the standard from $\theta \geq -1.05$. In the area of ability $\theta < -2.00$, where guessing is a decisive factor for examinees, DMAP typically loses to the standard Baysian methods, but there is not a large population in that ability area.

3. Estimation of ICC parameters

In this section we will demonstrate the implementation of the DMAP algorithm for obtaining 3PL ICC parameters on unknown (seeded) items, assuming that the ability of participating examinees has already been estimated. We will present our ICC functions, given by (1), as $P(a_i,b_i,c_i)(\theta)$; $i=1,\cdots,I$ to emphasize dependence on item parameters. There is a new $(I+1)$-th item with unknown parameters which is called a CAT seeded item; it is usually given to an examinee in the second, third, or fourth (random) position of his or her exam. If the CAT test is given to M examinees with abilities θ_m; $m=1,\ldots,M$, then the joint likelihood of the response vectors can be written as

$$L=\prod_{m=1}^{M} g(\theta_m)\cdot\prod_{i=1}^{K+1}(P(a_i,b_i,c_i)(\theta_m))^{u_{im}^*}\cdot(Q(a_i,b_i,c_i)(\theta_m))^{(1-u_{im}^*)}, \quad (6)$$

where $u_m^*=(u_{im}^*);i=1,\cdots,I+1$ is the binary vector of responses for examinee $m=1,\cdots,M$ on the test, including the seeded item. In expression (6) we took into account that the length of the test is increased to $(K+1)$ due to the inclusion of the seeded item. Relation (6) can be rewritten in the form:

$$L=L_0\cdot\prod_{m=1}^{M}(P(\bar{a},\bar{b},\bar{c}(\theta_m))^{\bar{u}_m^*}\cdot(Q(\bar{a},\bar{b},\bar{c})(\theta_m))^{(1-\bar{u}_m^*)}, \quad (7)$$

where $(\bar{a},\bar{b},\bar{c})=(a_{I+1},b_{I+1},c_{I+1})$ are the item parameters of the seeded item, and u_m^* is the response of m-th examinee on the seeded item in the test. Also L_0 here is the joint likelihood of the test without the seeded item.

To estimate item parameters for the seeded item, we must solve the problem of maximization of log-likelihood of (7), i.e., find a solution to the problem:

$$\ln L = \ln L_0 + \sum_{m=1}^{M} (\bar{u}_m^* \cdot \ln P(\bar{a},\bar{b},\bar{c})(\theta_m) + (1-\bar{u}_m^*) \cdot \ln(1 - P(\bar{a},\bar{b},\bar{c})(\theta_m)) \Rightarrow \max, \quad (8)$$

where $(\bar{a},\bar{b},\bar{c}) \in [a_{\min}, a_{\max}] * [b_{\min}, b_{\max}] * [c_{\min}, c_{\max}]$. The value of border segments such as $a_{\min}, a_{\max}, \ldots$ for different parameters are user-defined for a test, as in the case of ability estimation. We are interested in constrained maximization on the given parallelepiped-domain. The DMAP algorithm described below will check the border of this domain parallelepiped before going to the internal point. But if we assume the maximizing solution in (8) is reached on an inside point of the domain, we must find a solution of equalities:

$$\frac{\partial \ln L}{\partial a}(\hat{a},\hat{b},\hat{c}) = \frac{\partial \ln L}{\partial b}(\hat{a},\hat{b},\hat{c}) = \frac{\partial \ln L}{\partial c}(\hat{a},\hat{b},\hat{c}) = 0. \quad (9)$$

Then, from (9), we will have:

$$\frac{\partial \ln L}{\partial c} = \sum_{m=1}^{M} \left(\frac{u_m^*}{P(\bar{a},\bar{b},\bar{c})(\theta_m)} - \frac{1-u_m^*}{1-P(\bar{a},\bar{b},\bar{c})(\theta_m)} \right) \cdot \frac{\partial P}{\partial c}(\bar{a},\bar{b},\bar{c})(\theta_m).$$

However, from definition (1), the function $\frac{\partial P}{\partial c}(a,b,c)(\theta)$ does not depend on c. Using this fact, we have:

$$\frac{\partial^2 \ln L}{\partial^2 c} = \sum - \left(\frac{u_m^*}{P^2(\bar{a},\bar{b},\bar{c})(\theta_m)} + \frac{1-u_m^*}{(1-P(\bar{a},\bar{b},\bar{c})(\theta_m))^2} \right) \cdot \left(\frac{\partial P}{\partial c}(\bar{a},\bar{b},\bar{c})(\theta_m) \right)^2 < 0.$$

From this we can state that for fixed parameters $(\bar{a},\bar{b})$, the function $\ln L(\bar{a},\bar{b},\bar{c})$ is convex on c, and the function $\frac{\partial \ln L}{\partial c}(\bar{a},\bar{b},\bar{c})$ is monotone, decreasing on c. As in the case of estimation of ability, if $\ln L(\bar{a},\bar{b},\bar{c})$ is not reaching maximum on the border of the segment $[c_{\min}, c_{\max}]$, its

maximum is reached in the root of the function $\frac{\partial \ln L}{\partial c}(\bar{a},\bar{b},\bar{c})$ which can be found by a dichotomy process. Below, we describe in more detail how this work could be done in our case.

Let's introduce a function $F_m = 1+\exp(d\cdot a\cdot(\theta_m - b)); \quad m=1,\cdots,M$, then $\frac{\partial P(\bar{a},\bar{b},\bar{c})}{\partial c}(\theta_m)=\frac{1}{F_m}$, and $P(\bar{a},\bar{b},\bar{c})(\theta_m)=1+\frac{\bar{c}-1}{F_m}$. After some algebra we will have:

$$\frac{\partial \ln L}{\partial c}=\sum_{m=1}^{M}\left(\frac{u_m^{\bullet}}{F_m+\bar{c}-1}-\frac{1-u_m^{\bullet}}{1-\bar{c}}\right)=\sum_{m=1}^{M}\frac{u_m^{\bullet}}{F_m+\bar{c}-1}-\frac{N}{1-\bar{c}}. \qquad (10)$$

where N is the total number of wrong answers on the seeded item in the test. If $N=0$, i.e., there are no wrong answers, $u_m=1$, $m=1,\cdots,M$ for $c=1$ (case of "perfect guessing"), we will have $\frac{\partial \ln L}{\partial c}=\sum_{m=1}^{M}\frac{1}{F_m}>0$, which, due to monotone decreasing nature of function $\frac{\partial \ln L}{\partial c}$, means that $\frac{\partial \ln L}{\partial c}>0$ for all c, and so the log-likelihood function $\ln L$ is monotone, increasing function and reaching maximum on the right end $\bar{\bar{c}}=1$. If $N>0$ so there is examinee m_0 such that $u_{m_0}=0$, then $\frac{\partial \ln L}{\partial c}\to-\infty$ when $c\to 1$ and behavior of the function $\ln L$ depends on the behavior $\frac{\partial \ln L}{\partial c}$ on the left end $c=0$. If $c=0$; then

$$\frac{\partial \ln L}{\partial c}=\sum_{m=1}^{M}u_m^{\bullet}\frac{1}{F_m-1}-N=\sum_{m=1}^{M}u_m^{\bullet}\frac{F_m}{F_m-1}-M=\sum_{m=1}^{M}\frac{u_m^{\bullet}}{P_m}-M \qquad (11)$$

where $P_m = P(\bar{a},\bar{b},\bar{c})(\theta_m)$. From (11) it follows, if $\sum_{m=1}^{M}\frac{u_m^{\bullet}}{P_m}-M<0$, then the likelihood function is monotone, decreasing and reaching maximum on the left end $\bar{\bar{c}}=0$. If

$\sum_{m=1}^{M} \frac{u_m^*}{P_m} - M \geq 0$, we will have one root for function $\frac{\partial \ln L}{\partial c}(\bar{a}, \bar{b}, \bar{c})$ which can be found by dichotomy. This root $\bar{\bar{c}} = c(a,b)$ will provide the searching maximum likelihood. Utilizing this, we implement a search through the dense net of points $(\bar{a}_j, \bar{b}_j)$, $j = 1, \cdots, N$, where $(\bar{a}_j, \bar{b}_j) \in A \times B$, computing the likelihood $L(\bar{a}_j, \bar{b}_j, c(\bar{a}_j, \bar{b}_j))$ and getting approximate maximization, for which the precision depends on the density of the net. This search can be considerably decreased if we use a convexity of the function $L(\bar{a}, \bar{b}, c(\bar{a}, \bar{b}))$ on $\bar{b}$ for fixed $\bar{a} \in [a_{\min}, a_{\max}]$ (provided in the Appendix) under some approximation. Again, after more than 1,000,000 experiments, we can state that this approximation is holding in our case, i.e., the function $L(\bar{a}, \bar{b}, c(\bar{a}, \bar{b}))$ is convex on $\bar{b}$.

4. Comparing performances DMAP and BilogMG

Comparing the performance of the DMAP algorithm with the BilogMG algorithm is done through a set of simulations, but first the BilogMG package must be adjusted to get a reasonable performance. As we have explained, the matrix of responses for a CAT test is rather sparce. Further, items with low information are used very rarely, and items with high information are used too often, giving very non-uniform filling of the response matrix. As result, BilogMG very often leads to a non-convergence run, providing parameters too far from reality. To avoid this inconsistency, we run BilogMG in two stages. First, we simulate a paper-and-pencil test for our set of M examinees on items that belong to the CAT-ASVAB item pool. Then we run BilogMG

and save the result of item pool estimation with help of the BilogMG "SAVE" statement. After that, we run BilogMG for the data obtained from the simulated or real CAT-ASVAB test, with the seeded item included, using the preliminary estimation through the "IFNAME" subcommand in the "GLOBAL" statement in BilogMG. With this approach, BilogMG converges and provides a rather reasonable and stable estimation on the population of examinees with normal distributed abilities. After much experimentation, we are decided to use 30 quadrature points in the marginal estimations for BilogMG.

To compare performances in the "normal" situation, we use three typically representative items from the item pool for AR: an "easy" item $(a,b,c) = (1.17,-1.63,0.13)$, a "normal" item $(a,b,c) = (1.3,0.12,0.15)$, and a "hard" item $(a,b,c) = (1.23,1.63,0.07)$. All three items are rather informative in their areas of difficulty. Then, for each item we run the CAT-ASVAB test simulation twenty times for M examinees, changing random seeds each time to generate different response matrixes. In every run we use DMAP and adjusted BilogMG to re-estimate item parameters for the above described items. We found that both packages provide unbiased parameter estimation; the major differences are in the precision of those estimations.

First of all, we run our simulation for a different number of examinees with normal-normal distribution of their abilities, changing examinee number as:
$M \in \{300, 500, 750, 1000, 1500, 2000\}$. In this experiment, we try to identify the number of examinees needed to provide estimation of parameters which are most precise. In Table 1 we show estimation of SE for three parameters in our experiment.

(Table 1 about here.)

These results are graphically shown in the Figure 4 (Graphs depict variances of parameters estimation for a, b and c correspondingly).

(Figure 4 about here.)

As we can see, DMAP requires at least 1,000 examinees per test to get variances in a and b parameters comparable with BilogMG, and BilogMG is always better in the estimating of parameter c. However, the last advantage (more precise estimation of parameter c) disappears if we measure weighted average distances between "true" ICCs of studied items and ICCs built with estimated 3-PL parameters. Here, under weighted distance between two ICCs curves, we mean

$$D = \sqrt{\sum_{j=1}^{T} w_j \cdot (P(a,b,c)(\theta_j) - P(\tilde{a},\tilde{b},\tilde{c})(\theta_j)},$$

where $(\tilde{a},\tilde{b},\tilde{c})$ is the estimation of "true" parameters (a,b,c) by some package in a particular simulation experiment; θ_j, $j = 1,\ldots,50$ are equidistant points in ability domain $[-3.0, 3.0]$, and weights are normally distributed, i. e. $w_j \in N(0,1)$; $\sum_{j=1}^{T} w_j = 1$. In Table 2 and Figure 5 we show that, from the point of view of distances between ICC curves, both algorithms perform more or less equally, in spite of the fact that BilogMG approximates the guessing parameter c better than DMAP.

(Table 2 and Figure 5 about here).

This is because the influence of guessing parameter is strong where the density of the examinee population is small. From this simulation experiment, we see that the performance of both packages is about the same for $M = 1{,}500$, which we will assume in all further simulations;

we also begin calibration of a seeded item when the number of examinee answers on it is about 1,500 in "real" on-line calibration with CAT-ASVAB.

In the case of the CAT-ASVAB, very often we have violation of normality in examinee ability distribution due to seasonal and geographical location differences. To simulate this situation we consider two types of artificial populations. In the first type, we mix 750 examinees with normal-normal ability distribution with 750 examinees with ability distribution $N(-0.8, 1.0)$. After mixing, we get not-normal ability distributed population of examinees with mean of ability equal –0.4 and SE equal ≈ 1.15. We call this population "less able" (to the test). In the same mode, we make an "more able" population with mean +0.4 and the same SE ≈ 1.15. In both cases, we apply previously described simulation for the same three items of CAT-ASVAB Form 1 AR. We find the variances of estimation of 3-PL parameters are about the same as for the normal case (described above); the main differences are in biases of parameter estimations. Those biases are shown in Figure 6.

(Figure 6 about here.)

As we can see, BilogMG begins to be significantly biased in estimation of difficulty parameters, overestimates them for the "less able" population, and underestimates for the "more able" population. As a result, the average weighted distance between estimated ICCs and "true" ICCs significantly increases for BilogMG (Figure 7). On the other hand, the bias increases for DMAP are not significant with respect to the normal case.

(Figure 7 about here.)

As we have mentioned, there is one CAT-ASVAB test that is essentially not one-dimensional, General Science, which consists of three subtests: Physical Science, Biological

Science, and Chemical Science. To simulate the application of this test, we assume that every simulee has three abilities for every subtest, which are normal-normal distributed but highly correlated with a coefficient of correlation equal 0.8. Thus, the matrix of correlation for General Science abilities in this population looks like $R = \begin{pmatrix} 1.0 & 0.8 & 0.8 \\ 0.8 & 1.0 & 0.8 \\ 0.8 & 0.8 & 1.0 \end{pmatrix}$. We would like to get a three-dimensional ability vector $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3)$ such that every component of it will have a normal distribution with mean 0, and the correlation matrix between components will be equal R. To do this, we make a Cholesky decomposition of R, i.e., present it in the form $R = A^T * A$ where A^T matrix transposes to matrix A, the square root of R and $A = Q * diag(\sqrt{\lambda_i})$ where Q is a three-dimensional orthogonal matrix. In our case $\lambda_1 = \lambda_2 = 0.2$ and $\lambda_3 = 2.6$, and

$Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}$. Then, if vector $\bar{\theta} = (\theta_1, \theta_2, \theta_3)$ consists of three independent identically distributed components belonging to $N(0,1)$, vector $\tilde{\theta} = \bar{\theta} * A^T = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3)$ will have the desired multi-dimensional distribution (Bickel & Doksum, 1977). Thus, if a simulee gets a Physical Science item, we use $\tilde{\theta}_1$ ability to get the response for that item; if Biological, we use $\tilde{\theta}_2$; and if Chemical, we use $\tilde{\theta}_3$.

In this three-dimensional situation, we choose for simulation three representative items for each science: one "easy" item $b < -1.4$, one "normal" item $-0.3 < b < -0.3$, and one "hard" item $b > 1.7$ (altogether we choose nine items for the General Science test). As before, we run the simulation twenty times, changing random seeds and using 1,500 simulees in every run. Our

results show that both packages are not significantly biased in parameter estimation, but there are increases in variance estimation, compared with a one-dimensional test. These increases are shown in the Figure 8.

(Figure 8 about here.)

As we can see, the largest and most significant increase is in the variances of estimating difficulty parameters by BilogMG. Further, with BilogMG, we have a significant increase in weighted distance between the estimated and "true" ICCs, especially for "normal" items (Figure 9). On the other hand, the increase in the variances of estimating difficulty parameters by DMAP are not significant relative to the normal case.

(Figure 9 about here.)

5. Conclusion

We have demonstrated that the above described DMAP algorithm has about the same precision as the BilogMG algorithm in calibrating items from the CAT-ASVAB seeded design. More than that, in "special" circumstances, such as the absence of normality in prior distribution of examinee ability or the multi-dimensionality in item content, BilogMG loses its precision, but DMAP does not. This is because BilogMG is a marginal algorithm, with normality, to some extent, built in by the application of computation joint distribution through quadrature points. The other "weak" part of BilogMG is the application of only the Newton-Raphson algorithm as the main engine for local sub-optimization. As we have already mentioned, this tool will not pursue constrained optimization. However, from the point of view of maximization of joint likelihood, BilogMG and DMAP use different types of heuristics, so their solutions in different initial circumstances can be better or worse, depending on many "internal" conditions. Therefore, for real on-line calibration of CAT-ASVAB seeded items, we run both packages and choose the best solution by χ^2 evaluation.

Appendix

Convexity by Other Parameters

As we show, for fixed (a,b) the log-likelihood function $\ln L(a,b,c)$ is convex on c and reaches its maximum inside the prescribed segment $[c_{min}, c_{max}]$ or on its border. We now consider the case when the function $\ln L(a,b,c)$ reaches its maximum on c inside the above domain-segment. In this case there is a function $c = c(a,b)$ such that

$$\frac{\partial \ln L(a,b,c(a,b))}{\partial c} \equiv 0 . \qquad (12)$$

Because all considered functions are analytical under some regularity conditions (Kantorovich, 1968), the function $c = c(a,b)$ is also analytical, so it has all the derivatives. Let us present our 3PL function in the form:

$$P(a,b,c)(\theta) = c + (1-c)\cdot P_0(a,b)(\theta), \qquad (13)$$

where $P_0(a,b)(\vartheta) = \frac{\exp(L(a,b,\theta))}{1+\exp(L(a,b,\theta))}$, i.e., $P_0(a,b)(\vartheta)$ is a 2PL ICC in the considered case (Here $L(a,b,\theta) = D\cdot a\cdot(\theta - b)$). Using (13) we can rewrite identity (12) in the form:

$$\frac{\partial \ln L(a,b,c(a,b))}{\partial c} \equiv \sum_{m=1}^{M} \left(\frac{u_m^*}{P(a,b,c(a,b))(\theta_m)} - \frac{1-u_m^*}{1-P(a,b,c(a,b))(\theta_m)}\right)(1 - P_0(a,b)(\theta_m)) . \qquad (14)$$

Then for the derivative of $\ln L(a,b,c(a,b))$ with respect to b we have:

$$\frac{\partial \ln L(a,b,c(a,b))}{\partial b} = \sum_{m=1}^{M} \left(\frac{u_m^*}{P(a,b,c(a,b))(\theta_m)} - \frac{1-u_m^*}{1-P(a,b,c(a,b))(\theta_m)}\right)\cdot \frac{\partial P(a,b,c(a,b))(\theta_m)}{\partial b} \quad \text{and}$$

$$\frac{\partial^2 \ln L(a,b,c(a,b))}{\partial b^2} = \sum_{m=1}^{M} -\left(\frac{u_m^*}{(P(a,b,c(a,b))(\theta_m))^2} + \frac{1-u_m^*}{(1-P(a,b,c(a,b))(\theta_m))^2}\right)\cdot\left(\frac{\partial P(a,b,c(a,b))(\theta_m)}{\partial b}\right)^2$$

$$+\sum_{m=1}^{M}\left(\frac{u_m^*}{P(a,b,c(a,b))(\theta_m)} - \frac{1-u_m^*}{1-P(a,b,c(a,b))(\theta_m)}\right)\cdot\frac{\partial^2 P(a,b,c(a,b))(\theta_m)}{\partial b^2}$$

The first sum in this expression has a negative value. To work with the second sum, let us consider the expression for the second derivative $\frac{\partial^2 P(a,b,c(a,b))(\theta)}{\partial b^2}$. Taking a derivative of (13) we have:

$$\frac{\partial P(a,b,c(a,b))(\theta)}{\partial b} = \frac{\partial c(a,b)}{\partial b}\cdot(1-P_0(a,b)(\theta)) - (1-c)\cdot\frac{D\cdot a}{(1+\exp(L(a,b,\theta)))^2}.$$

From which expression we get:

$$\frac{\partial^2 P(a,b,c(a,b))(\theta)}{\partial b^2} = \frac{\partial^2 c(a,b)}{\partial b^2}\cdot(1-P_0(a,b)(\theta)) + 2\cdot\frac{\partial c(a,b)}{\partial b}\cdot\frac{D\cdot a}{(1+\exp(L(a,b,\theta)))^2}$$

$$+2\cdot(1-c(a,b))\cdot\frac{D^2\cdot a^2}{(1+\exp(L(a,b,\theta)))^3} \qquad (15)$$

Here we utilize expression $\frac{\partial P_0(a,b)(\theta)}{\partial b} = \frac{D\cdot a}{(1+\exp(L(a,b,\theta)))^2}$, taking into account that $\frac{1}{(1+\exp(L(a,b,\theta)))^n} = (1-P_0(a,b)(\theta))^n$. Then, from (5) we will have:

$\frac{\partial^2 P(a,b,c(a,b))(\theta)}{\partial b^2} \approx \frac{\partial^2 c(a,b)}{\partial b^2}\cdot(1-P_0(a,b)(\theta))$ which, together with identity (14), will get us to the conclusion that the type of approximation $\frac{\partial^2 L(a,b,c(a,b))}{\partial b^2}$ is negative, so the function $L(a,b,c(a,b))$ is convex on b for fixed a. The same type of consideration can be given about convexity of $L(a,b,c(a,b))$ with respect to a for fixed b.

References

Baker, F. D. (1992). Item response theory. New York: Marcel Dekker Inc.

Bickel, P. J., & Doksum, K. A. (1977). Mathematical Statistics. San Francisco, CA: Holden-Day, Inc.

Blum, E. K. (1972). Numerical analysis and computation: Theory and Practice. Reading, MA: Addison Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, Psychometrika, 46, #4, 443-458.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized Adaptive Testing (pp. 141-145). Washington, DC: American Psychological Association.

Kantorovich, L. V. (1968). Functional analysis and applied mathematics. Washington, DC: NBS.

Krass, I. A. (1997, June). Getting more precision on computer adaptive testing. Paper presented at the 62nd Annual meeting of Psychometric Society, University of Tennessee, Knoxville, TN.

Levine, M. (1984). An introduction to multilinear formula scoring theory. (Office of Naval Research Report 84-4). Champaign, IL: University of Chicago.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

McLachlan, G. J., & Krishnan, T. (1997). The EM algorithm and Extensions. New York: John Wiley & Sons.

Owen, R. J. (1975). A Baysian sequential procedure for quantal response in the context of adaptive mental testing. Journal of American Statistical Association, 70, 351-356.

Ramsay, J. O. (1975). Solving implicit equations in psychometric data analysis, Psychometrika, 40, 337-360.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 38, 221-233.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized Adaptive Testing (pp. 131-140). Washington, D C: American Psychological Association.

Thissen, D., & Steinberg. L. (1984). A model for multiple choice items. Psychometrika, 49, 501-519.

Zimowski, M. F., & Bock, R. D. (1987). Full-information item factor analysis from the ASVAB CAT pool. (Methodology Research Center Report #87-1), Chicago: University of Chicago.

Table 1. Variances of 3PL parameters in the "Normal" simulation

	A-parameter		B-parameter		C-parameter	
	BLG	DMAP	BLG	DMAP	BLG	DMAP
2000	0.0378	0.022	0.023	0.0154	0.0007	0.0041
1500	0.0231	0.0395	0.0231	0.0122	0.0004	0.0043
1000	0.0308	0.0428	0.0237	0.0169	0.0005	0.0051
750	0.0342	0.0428	0.025	0.0262	0.0006	0.0067
500	0.0668	0.0923	0.0318	0.0246	0.0003	0.0072
300	0.0942	0.2462	0.0428	0.0419	0.0006	0.0071

Table 2. Average distances between ICCs

	BLO	DMAP
2000	0.0221	0.0185
1500	0.0227	0.0196
1000	0.0235	0.0247
750	0.0254	0.0258
500	0.0322	0.0292
300	0.0399	0.0416

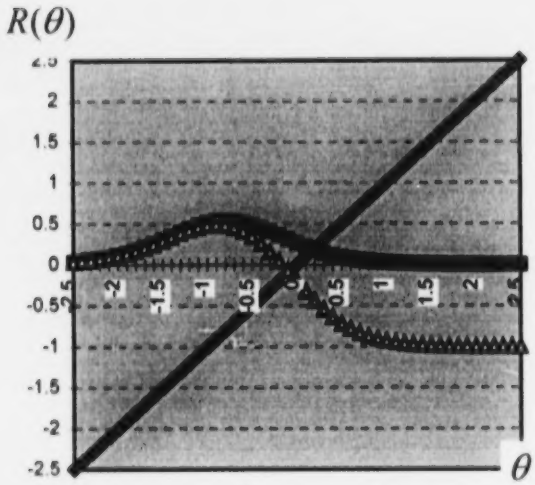

FIGURE 1.

The case of a test of length two, where the first item was answered correctly and the second wrongly.

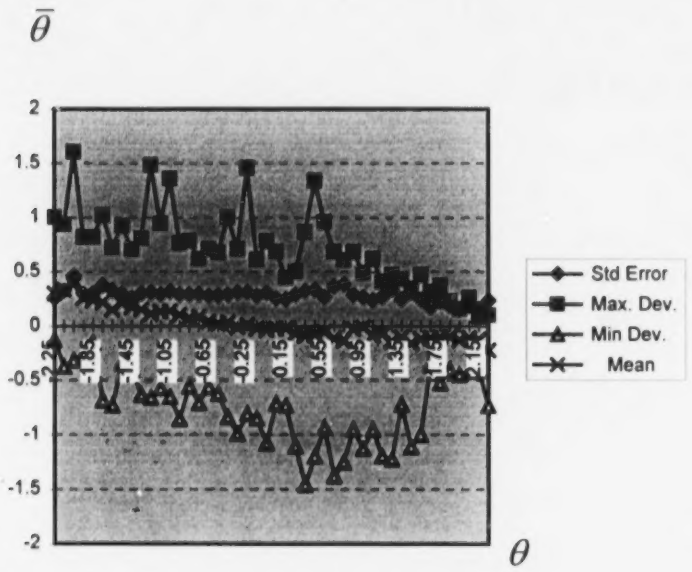

FIGURE 2

Results of AR simulation after standard Baysian implementation.

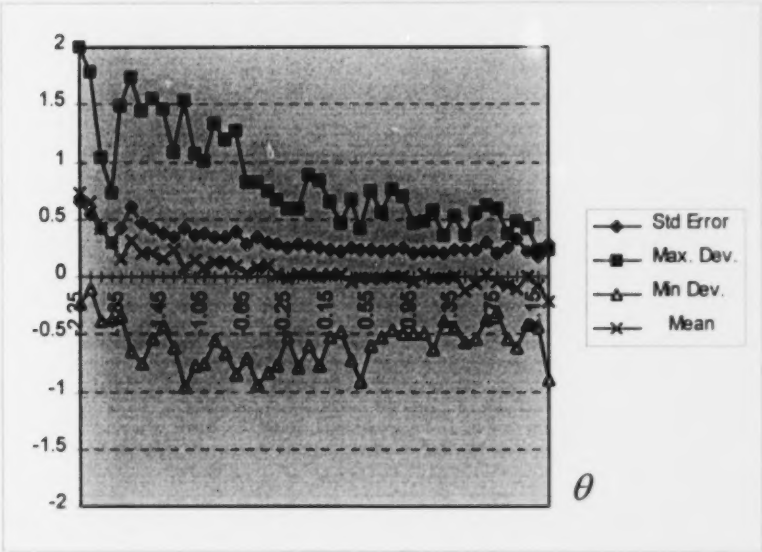

FIGURE 3.

Results of AR simulation after DMAP implementation.

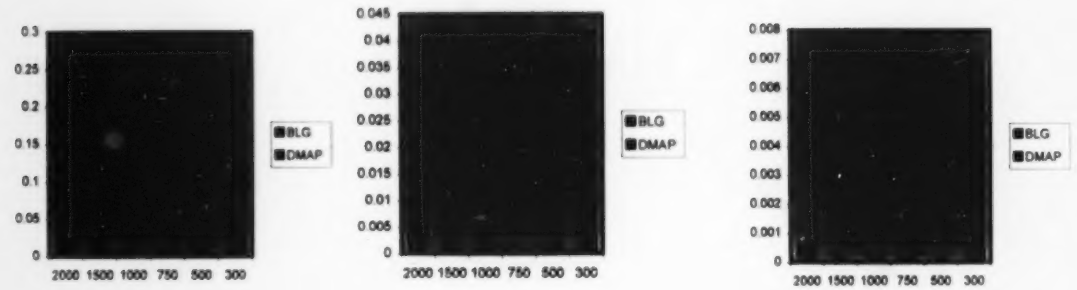

FIGURE 4.

Variances of 3-PL parameters in the "Normal" simulation.

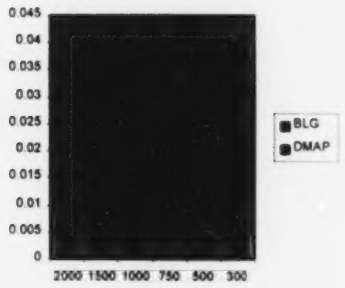

FIGURE 5.

Average distances between ICCs.

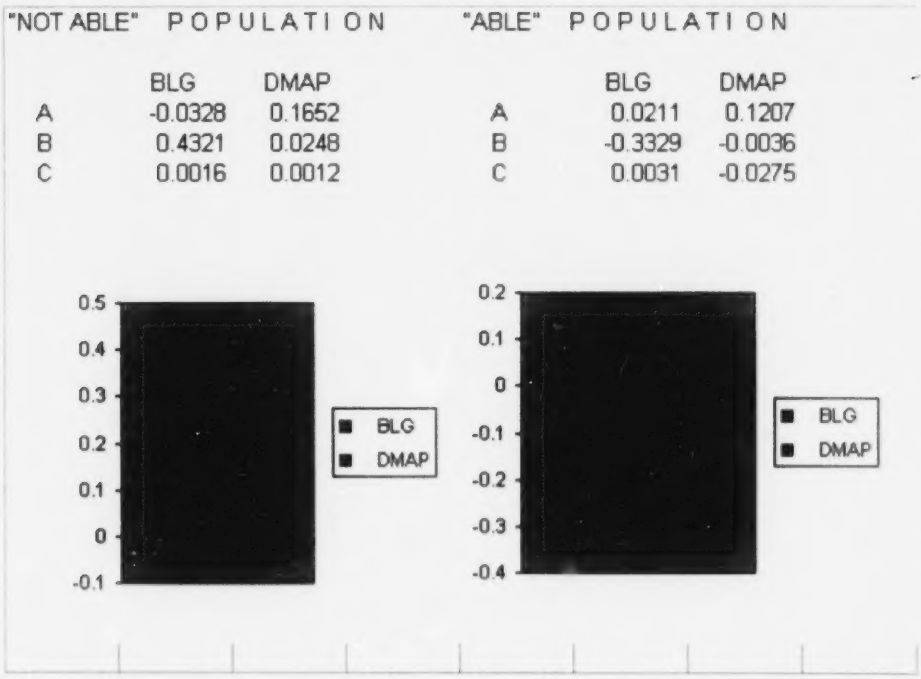

FIGURE 6.

Biases in the case of not "Normal" population.

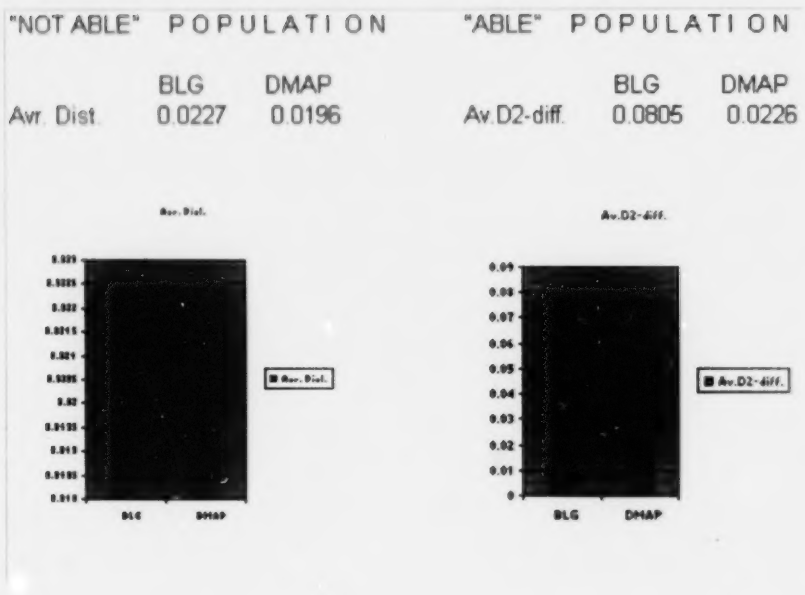

FIGURE 7.

Weighted ICCs differences in the case of not "Normal" population.

	BLG	DMAP
Incr Vr. A.	0.0461	0.0325
Incr Vr. B.	0.1392	0.0134
Incr Vr. C.	0.0005	0.0038

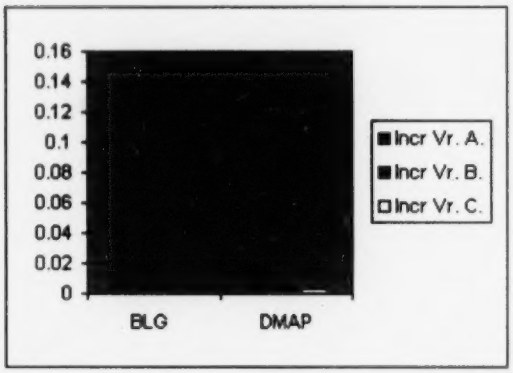

FIGURE 8.

Increases of variances in three-dimensional case.

	BLG	DMAP
Easy	0.0129	-0.001
Normal	0.0336	0.0122
Hard	0.0119	0.0066
Overall	0.0194	0.006

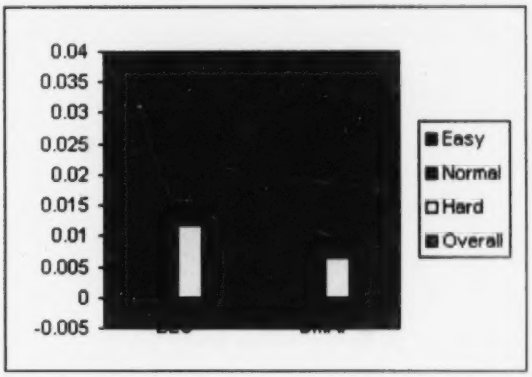

FIGURE 9.

Increase in distances between ICCs.